Agilent GeneSpring Workgroup 5.3 Web Application

User Guide

This manual provides information on the use of GeneSpring Workgroup 5.3 Web Application for mining and analysis of microarray data within Workgroup Server.

Getting Started 2

Analysis Workflow 6

Searching and Inspecting Data 20

Gene Selection 34

Data Export 37

Statistical References 39

What is GeneSpring Workgroup Web Application?

Workgroup Web Application is an intuitive user interface to gene expression data stored within GeneSpring Workgroup Server. Web Application allows non-expert users to search, filter, and perform exploratory analysis on microarray profiles from within a secure, centralized environment. With Workgroup Web Application, you can:

- Apply multiple filter conditions to effectively search for and combine samples of interest and filter for differentially expressed genes based upon p-value and fold-change
- Analyze sample trends using principle component analysis.
- Determine differential expression groupings using heatmap visualization.
- Produce expression profiles for genes of interest across samples.

Agilent Technologies

Getting Started

Setting up a client PC

Workgroup Web Application runs from within a web browser. The Web Application makes use of client-server transactions with cookies and JavaScript active scripting.

Browser

The Web Application has been tested on Firefox 1.5 and Internet Explorer 6.0 SP2 and greater. To check the version of your web browser:

1 Open Browser.

2 Select **Help > About Browser**.

Active scripting

Enable JavaScript (Active Scripting) if necessary.

To do this in Firefox 1.5 and greater:

1 Open Browser.

2 Select **Tools > Options...**

3 In the **Options** dialog, click the **Content** icon.

4 Select the **Enable JavaScript** option.

5 Click **OK**.

To do this in Internet Explorer 6.0 and greater:

1 Open Browser.

2 Select **Tools > Internet Options...**

3 In the **Internet Options** dialog, click the **Security** tab.

4 Click the **Custom Level...** button.

5 In the **Security Settings** dialog, scroll down until you see the **Scripting** section.

6 Under **Active Scripting**, select the **Enable** option.

7 Click **OK**.

Cookies

Web Application uses pieces of information stored on your computer called cookies to remember display preferences.

CAUTION Allowing cookies or changing privacy settings can have undesirable consequences when visiting untrusted sites on the internet. See your system administrator for the preferred way of allowing cookies from the Web Application at your site.

To enable cookies in Firefox 1.5 and greater:

1 Open Browser.

2 In the **Options** dialog, click the **Privacy** icon.

3 Click the **Cookies** tab.

4 Select **Allow sites to set cookies**.

5 Click **OK**.

To enable cookies in Internet Explorer 6.0 and greater:

1 Open Browser.

2 Select **Tools > Internet Options...**

3 Click the **Privacy** tab.

4 Examine privacy settings level. If privacy settings are set above Medium, reduce settings to Medium to allow automatic handling of cookies. Alternatively, click the Advanced tab.

5 In the **Advanced Privacy Settings** dialog, select the **Override automatic cookie handling** check box. Select the accept radio button for both First-party cookies and Third-party cookies. Select the Always allow session cookies check box.

6 Click **OK**.

Starting a Web Application Session

GeneSpring Workgroup Web Application is an interactive data mining and analysis interface to an existing GeneSpring Workgroup 5.3 Server. The interface holds current workflow selection and parameter information in memory, called a session. The session is also responsible for collecting samples of interest for subsequent analysis steps within a workflow. The server can manage many simultaneous, independent secure sessions.

NOTE

See your system administrator for the correct Uniform Resource Locator (URL) of the Workgroup Server and for your secure login credentials. URLs are commonly known as an “internet address” and begin with the protocol indicator “http”. URLs can be actual addresses, as in “http://192.168.0.1” or names as in “http://www.myserver.myco”.

Log In

To begin a Web Application session, you must login to a GeneSpring Workgroup Server:

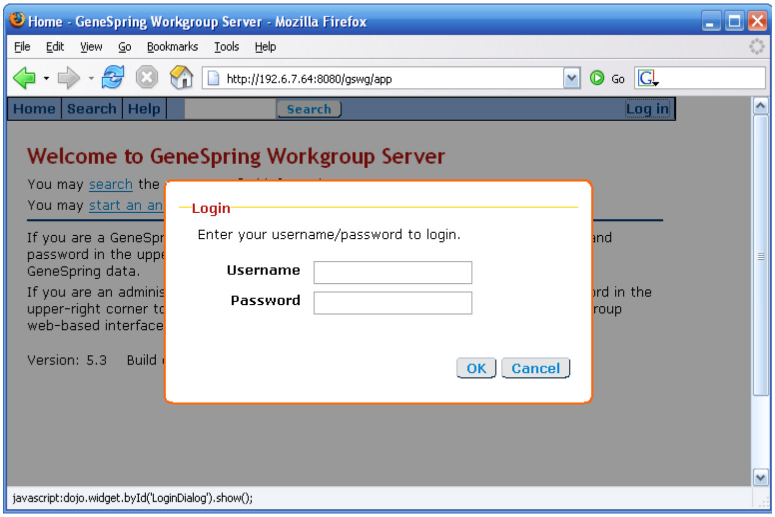
1 Open Browser.

2 In the **Navigation Textbox** of the browser, enter the URL of the Workgroup Server.

3 In the Web Application webform, click **Log In**.

4 In the Login dialog box, enter your Username and Password.

5 Click **OK** to begin a Web Application session, or **Cancel** to abort.


Home - GeneSpring Workgroup Server - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://192.6.7.64:8080/gswg/app
Go
Home
Search
Help
Search
Log in
OK
Cancel

Welcome to GeneSpring Workgroup Server

You may search the

You may start an an

If you are a GeneSpr
password in the uppe
GeneSpring data.

If you are an adminis
upper-right corner to
web-based interface

Version: 5.3 Build

Login

Enter your username/password to login.

Username

Password

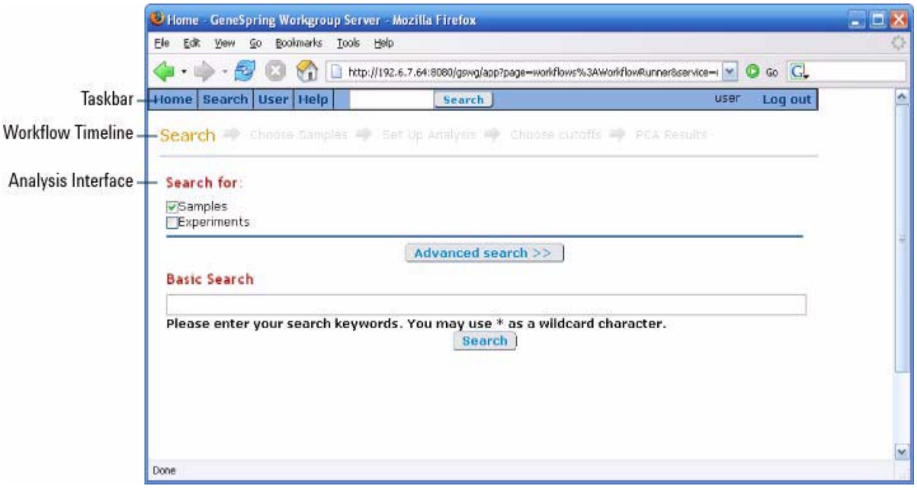
javascript:dojo.widget.byId('LoginDialog').show();

Figure 1 The Login dialog box.

Analysis Workflow

Workflow Display

The main display for analysis within Workgroup Web Application is centered around a workflow timeline. The timeline reflects the current status of an interactive session. Each analysis workflow is divided into several interconnected links to webforms that collect information about the data to be analyzed.



Please enter your search keywords. You may use * as a wildcard character.

Figure 2 Main analysis workflow display.

The workflow display consists of three basic components:

Taskbar

The Taskbar contains several quick links to pages within the Web Application and is meant to supplant use of the native navigation buttons within the browser. It is accessible from all pages in the Web Application.

Home Navigates to the entry page of the Web Application. This is the top of a hierarchical analysis workflow. It is necessary to start at the **Home** page whenever you would like to switch between workflows.

Search **Search > Simple/Advanced Search** will bring you to the general search function of the Web Application. This link is for data mining only and does not link to any workflow analysis. If you would like to start an analysis workflow, select Home.

Help Displays the online *User Guide*.

Search Box Calls the basic search function from Web Application (see "Basic Search" on page 20).

Log In Name Displays the username currently logged into the Web Application, or blank if no user is logged in.

Log In / Log Out Toggles between login states on the Workgroup Server. If the Log In link is displayed, the user is not currently logged into Workgroup Server and clicking the link launches the login dialog box (see Figure 1 on page 5). If the Log Out link is displayed, the user is logged into Workgroup Server and clicking on the link will log the user out of Workgroup Server. It is necessary to login prior to analysis with the Web Application.

Workflow Timeline

A workflow in the Web Application is tracked in memory for easy navigation by the Workflow Timeline. The orange, boldfaced entry corresponds to the currently active workflow step. Each blue colored entry represents a link to a completed step in the analysis workflow. Grey entries show workflow steps not yet completed. The workflow timeline is accessible after an analysis workflow has been chosen.

Analysis Interface

The bulk of the display contains webforms for setting up and completing an analysis. This is also the place where visualizations of workflow analysis will be presented. The analysis interface is dynamic in nature, reflecting the current Web Application task.

Available Workflows

Analysis within the Web Application consists of parallel workflows that share common conditional access to data stored in the Workgroup Server.

All analysis workflows begin with a search for samples stored in the Workgroup database using sample attributes and annotations. Based on search criteria, a list of samples is returned, grouped by array type. Analysis is available for a subset or all of the returned samples with the requirement that selected samples share a single array type.

A virtual GeneSpring experiment is created from the selected samples. This experiment and any objects created for it exist for the length of a single search session. Following sample selection, navigation within the workflow (using the workflow timeline) is possible. The same sample set will be applied unless new search criteria are applied or a different set of samples is chosen.

Comparisons across samples grouped by an existing parameter or attribute may be performed.

There are three available workflows:

Sample Comparison (PCA)

This workflow will allow you to compare the gene expression profiles of selected samples using Principle Component Analysis (PCA). You will have the option to first identify genes that are differentially expressed between selected experimental conditions, then compare the expression of those genes between the samples representing those conditions. The PCA scores for the first three components will be reported in an exportable table and presented with visualizations of those components. For more information on PCA analysis, see “Principle Component Analysis” on page 42.

Differential Expression with Heatmap

This workflow will allow you to identify genes that are differentially expressed between experimental conditions. Gene expression fold-change and p-value thresholds may be applied to generate a two dimensional heatmap formed by the expression value intersect of two dendrograms. In one dendrogram, genes are grouped according to the degree of similarity between experimental condition expression profiles. In the other dendrogram, samples are grouped according to the degree of similarity between the gene expression profiles. For more information on clustering and Heatmap analysis, see "Hierarchical Clustering" on page 41.

Gene Expression Profile(s) This workflow will allow you to look at the expression level of genes of interest across selected samples in a line graph view. The normalized intensity values for each gene in each sample will be reported in an exportable table.

The following diagram shows the activities involved in analyzing data within the Web Application.

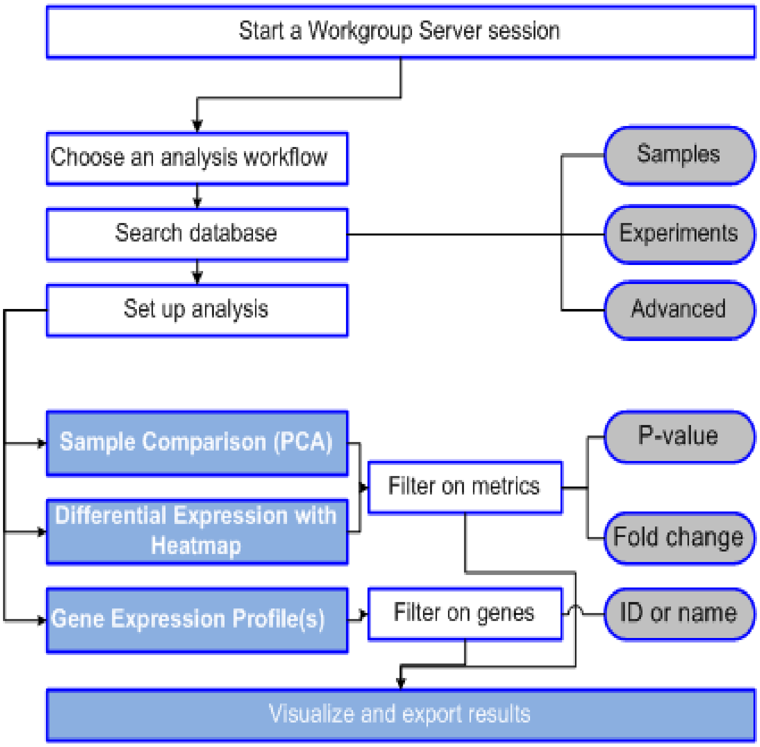

Start a Workgroup Server session
Choose an analysis workflow
Search database
Set up analysis
Samples
Experiments
Advanced
Sample Comparison (PCA)
Differential Expression with Heatmap
Gene Expression Profile(s)
Filter on metrics
Filter on genes
P-value
Fold change
ID or name
Visualize and export results

Figure 3 Workflow roadmap. Selecting samples from a database search initiates a session based upon those samples.

Sample Comparison (PCA)

The biological response to various experimental conditions can be characterized, in part, by the gene expression profile that is induced. For example, suppose a compound induces a certain gene expression profile and has been shown to have tumor suppressive activity. Compounds that induce a similar gene expression profile may potentially possess tumor suppressive qualities. Thus, it is often of interest to compare the gene expression profiles of biological samples exposed to different experimental conditions. A useful visualization tool for such analysis is Principle Component Analysis, or PCA.

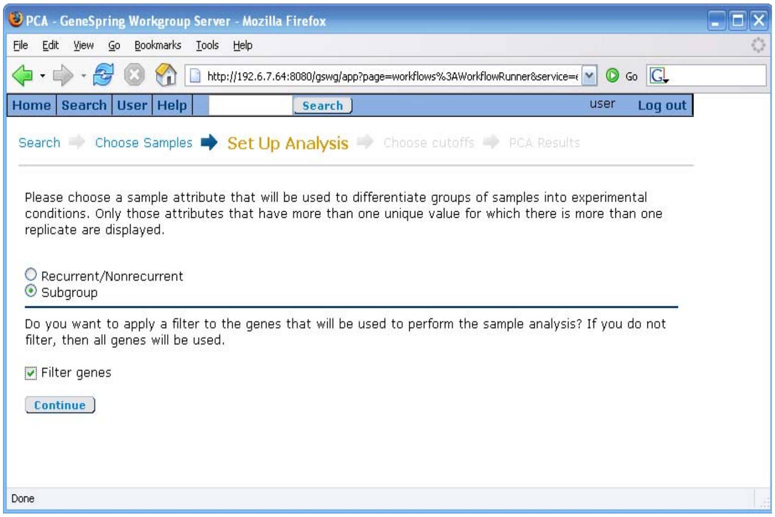
PCA is a mathematical procedure used to reduce the dimensionality of complex datasets by transforming correlated variables in the data into fewer uncorrelated variables while retaining as much information as possible. In doing so, PCA identifies patterns in the data and expresses data in a way that highlights the similarities and differences between the variables. Performing PCA on samples will uncover similarities and differences in the expression profiles of the selected samples.

The following procedure outlines the steps used to analyze gene expression profiles of samples using PCA. Window references follow a typical workflow analysis using the data set from “Gene expression correlates of clinical prostate cancer behavior.” Singh D, et. al., Cancer Cell. 2002 Mar;1(2):203-9.

1. In the Taskbar, select **Home**.
2. In the Home Page, click on the **Start An Analysis** link.
3. In the Choose An Analysis Workflow page, click on the **Sample Comparison (PCA)** link.
4. Search for samples using keywords or attributes (see “Searching and Inspecting Data” on page 20).
5. Select samples for analysis by checking the boxes in the first column of the tabular list that correspond to samples of interest (see “Data Selection” on page 27).
6. Click **Choose Samples** to analyze the samples selected in the previous step.

7 In the Set Up Analysis webform, select a comparison based on available attributes that will be used to differentiate groups of samples into experimental conditions. Only those attributes that have more than one unique value and for which there is more than one replicate are displayed.

Using the example data set, select a differentiation attribute called subgroup, which corresponds to the surgical margin recorded following patient prostatectomy. Note that the Filter Genes option is selected by default.


PCA - GeneSpring Workgroup Server - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.6.7.64:8080/gswg/app?page=workflows%3AWorkflowRunner&service=
Go
Home Search User Help
Search
user
Log out
Search
Choose Samples
Set Up Analysis
Choose cutoffs
PCA Results
Done

Please choose a sample attribute that will be used to differentiate groups of samples into experimental conditions. Only those attributes that have more than one unique value for which there is more than one replicate are displayed.

○ Recurrent/Nonrecurrent
● Subgroup

Do you want to apply a filter to the genes that will be used to perform the sample analysis? If you do not filter, then all genes will be used.

☑ Filter genes

Continue

Figure 4 Set Up Analysis webform

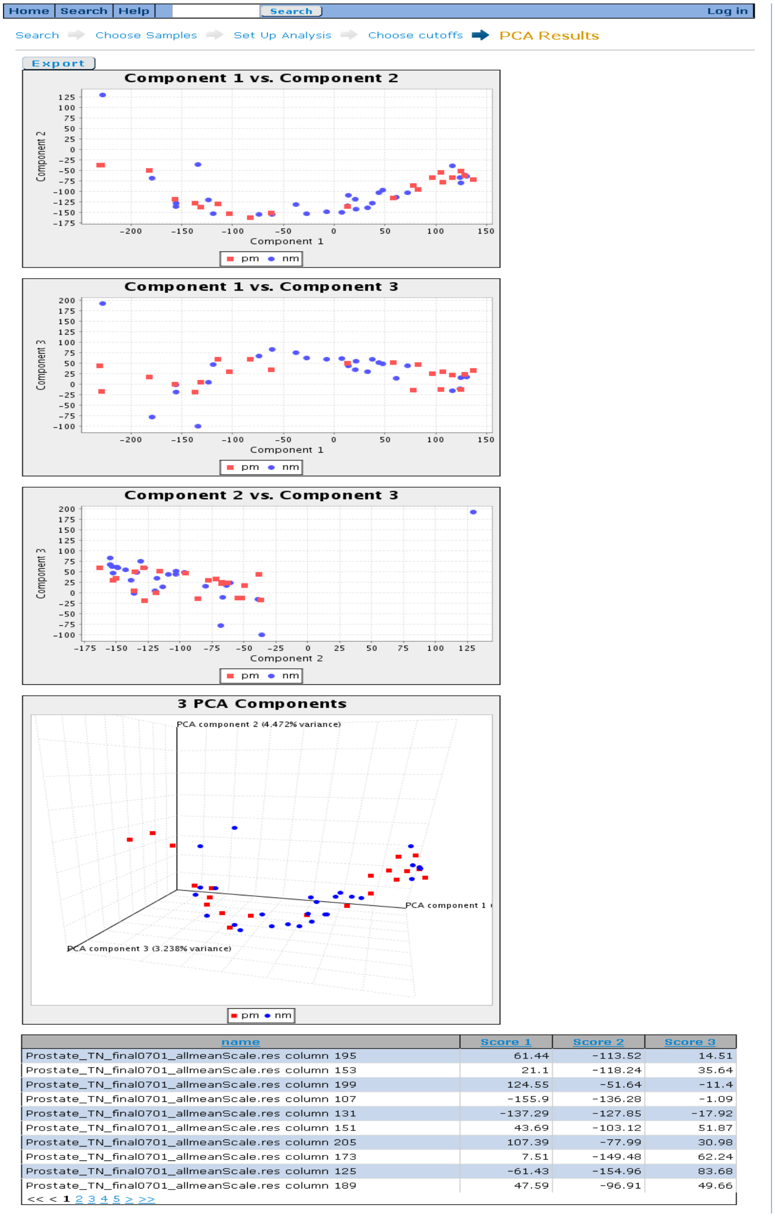
8 The Filter Genes option is checked by default. Deselect the checkbox if all genes are to be used in the PCA analysis. Click **Continue**.

9 If you wish to Filter Genes prior to PCA analysis, select p-value and fold-change cutoffs and the baseline condition from the Choose Cutoffs webform (see “Choose Cutoffs for Gene Filtering Based on Scores” on page 34).

10 Click **Update** to see how many genes pass the proposed filters from the previous step. Click **Continue** when satisfied with the filter, or return to step 9 and adjust the filter cutoff levels.

11 Analyze and export component visualizations and tabular data (see Figure 5 on page 12 and “Data Export” on page 37).

In this example an interesting trend is visible between conditions pm (positive surgical margin, meaning there was evidence of tumor cells at the periphery of the excised prostate tissue) and nm (negative surgical margin). Positive surgical margins tend to lie at the extreme PCA values along the first principle component (eigenvector). Further analysis and classification is achieved through examination of the tabular output of the first three principle component scores.



name	Score 1	Score 2	Score 3
Prostate_TN_final0701_allmeanScale.res column 195	61.44	-113.52	14.51
Prostate_TN_final0701_allmeanScale.res column 153	21.1	-118.24	35.64
Prostate_TN_final0701_allmeanScale.res column 199	124.55	-51.64	-11.4
Prostate_TN_final0701_allmeanScale.res column 107	-155.9	-136.28	-1.09
Prostate_TN_final0701_allmeanScale.res column 131	-137.29	-127.85	-17.92
Prostate_TN_final0701_allmeanScale.res column 151	43.69	-103.12	51.87
Prostate_TN_final0701_allmeanScale.res column 205	107.39	-77.99	30.98
Prostate_TN_final0701_allmeanScale.res column 173	7.51	-149.48	62.24
Prostate_TN_final0701_allmeanScale.res column 125	-61.43	-154.96	83.68
Prostate_TN_final0701_allmeanScale.res column 189	47.59	-96.91	49.66

<< < 1 2 3 4 5 > >>

Figure 5 PCA results: visualization and tabular view.

Differential Expression with Heatmap

Identification of genes that are differentially expressed between a set of conditions is often the first step in an attempt to understand the biological process under examination. Suppose that you are comparing normal human breast and human breast carcinoma samples. One hypothesis states that the difference in phenotype may have resulted from differential expression of certain genes. Isolating these genes may lead to the identification of key genes that mediate the underlying mechanism of the disease. It may also aid in the identification of potential therapeutic drug targets.

Once these genes of interest have been identified, it is often of interest to cluster genes with similar expression profiles. Functionally related genes have been shown to share similar expression profiles. Thus, the biological function of previously uncharacterized genes may be deduced by determining which genes are closely clustered with them within a hierarchical dendrogram. Additionally, genes that share similar expression profiles may be co-regulated, allowing for information about the regulatory systems of these genes to be delineated.

Clustering samples on the similarity of the expression of genes of interest can be used to assess similarity of expression profiles of different tissues, cancer sub-types, or in the case of drug compounds, the relative effect of those compounds on gene expression.

In hierarchical clustering of genes, genes are ordered in a dendrogram according to the degree of similarity between their expression profiles across a set of conditions. In other words, genes with similar expression profiles across a set of conditions will be grouped closer together in the dendrogram than genes with less similar expression profiles. The dendrogram that results from hierachical clustering of genes will be displayed vertically. The same technique may be applied to samples, where the samples are ordered in the dendrogram according to the degree of similarity between the expression of their genes. The dendrogram that results from hierachical clustering of samples will be displayed horizontally.

A useful visualization technique is to simultaneously display color-coded results from hierachical clustering of genes and hierarchical clustering of samples. Such a two-dimensional clustering approach yields a matrix with entry values equal to the expression index of a particular gene in a given sample. Color representation of such a matrix is commonly known as a “heatmap”.

The following outlines the steps to analyze differential gene expressions of samples with Heatmap. Window references follow a typical workflow analysis using the data set from “Gene expression correlates of clinical prostate cancer behavior.” Singh D, et. al., Cancer Cell. 2002 Mar;1(2):203-9.

1 In the Taskbar, select **Home**.

2 In the Home Page, click on the **Start An Analysis** link.

3 In the Choose An Analysis Workflow page, click on the **Differential Expression with Heatmap** link.

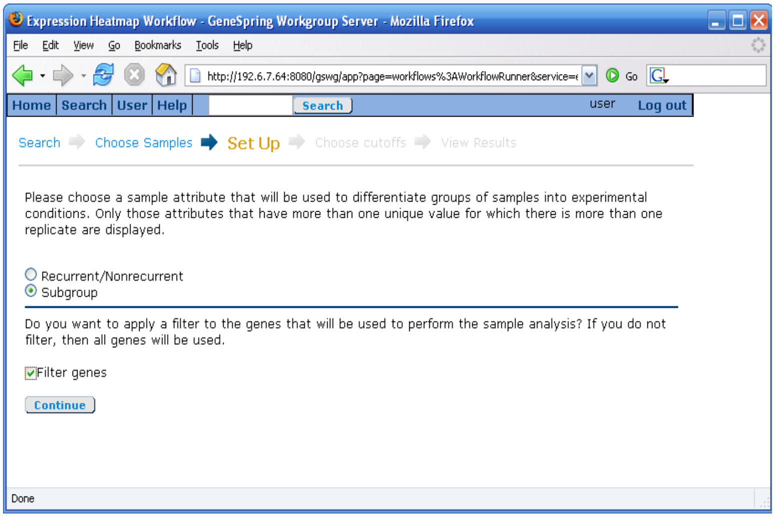
4 Search for samples using keywords or attributes (see “Searching and Inspecting Data” on page 20).

5 Select samples for analysis by checking the boxes in the first column of the tabular list that correspond to genes of interest (see “Data Selection” on page 27).

6 Click **Choose Samples** to analyze the samples selected in the previous step.

7 In the Set Up Analysis webform, select a comparison based on available attributes that will be used to differentiate groups of samples into experimental conditions. Only those attributes that have more than one unique value and for which there is more than one replicate are displayed.

Using the example data set, select a differentiation attribute called subgroup, which corresponds to the surgical margin recorded following patient prostatectomy. Note that the Filter Genes option is checked by default.


Expression Heatmap Workflow - GeneSpring Workgroup Server - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.6.7.64:8080/gswg/app?page=workflows%3AWorkflowRunner&service=
Go
Home
Search
User
Help
Search
user
Log out
Search
Choose Samples
Set Up
Choose cutoffs
View Results
Continue
Done

Please choose a sample attribute that will be used to differentiate groups of samples into experimental conditions. Only those attributes that have more than one unique value for which there is more than one replicate are displayed.

○ Recurrent/Nonrecurrent
● Subgroup

Do you want to apply a filter to the genes that will be used to perform the sample analysis? If you do not filter, then all genes will be used.

☑ Filter genes

Figure 6 Set Up Analysis webform

8 The Filter Genes option is checked by default. Deselect the checkbox if all genes are to be used in the Heatmap generation. Click on continue.

9 If you wish to Filter Genes prior to Heatmap analysis, select p-value and fold-change cutoffs and the baseline condition from the Choose Cutoffs webform (see “Choose Cutoffs for Gene Filtering Based on Scores” on page 34).

NOTE

Generation of the Heatmap may take a noticeable amount of time, during which the Web Application will display a Refresh button within the Analysis Interface. Click Refresh to display the Heatmap.

10 Analyze and export the Heatmap visualization (see Figure 7 on page 16 and “Data Export” on page 37).

In this example, notice several significant clusters of similar gene expression across samples delineated by conditions pm (positive surgical margin, meaning there was evidence of tumor cells at the periphery of the excised prostate tissue) and nm (negative surgical margin).

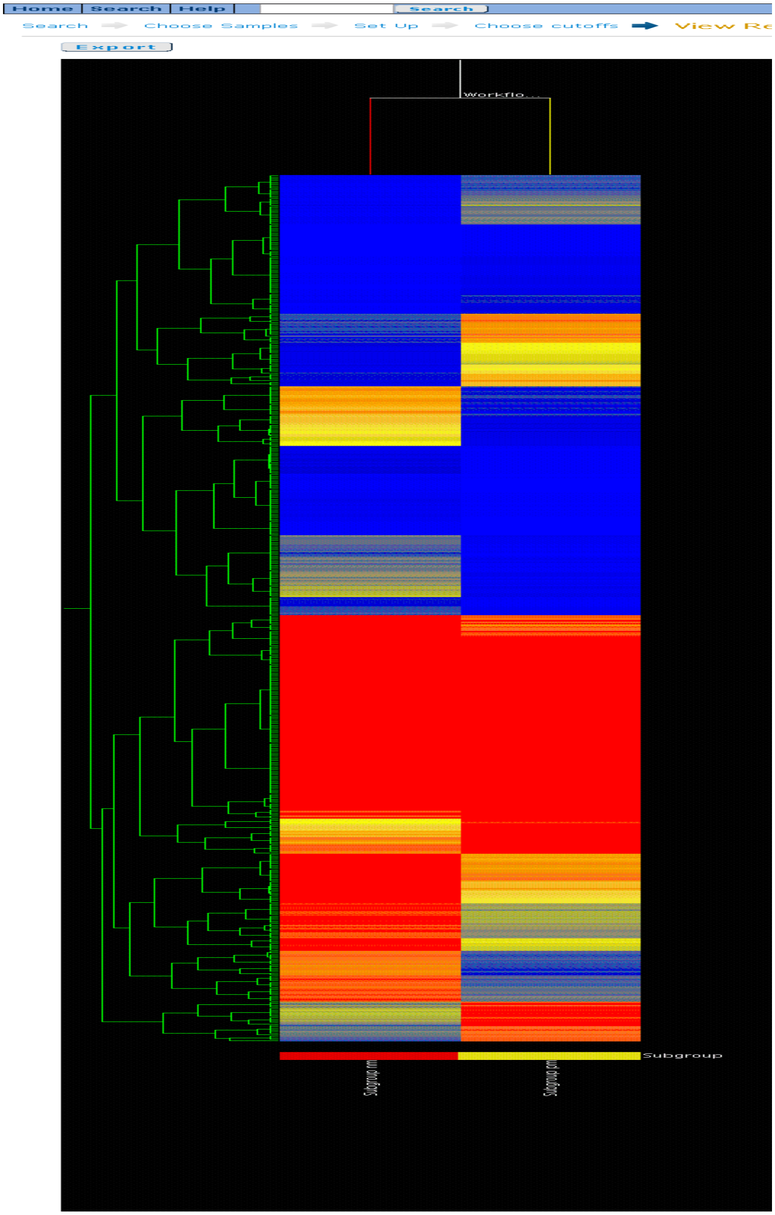


Figure 7 Heatmap results: visualization.

Gene Expression Profile(s)

It is often of interest to view how the expression of a number of genes of interest changes across samples from different experimental conditions such as treatment types, tissue types, or time points. A useful visualization technique is to plot the expression value of a given gene across a number of samples selected during data mining in a line graph. Such a simplified representation makes it easy to spot similar trends in the gene expression profiles under comparison.

The following outlines the steps to analyze gene expression profile(s). Window references follow a typical workflow analysis using the data set from “Gene expression correlates of clinical prostate cancer behavior.” Singh D, et. al., Cancer Cell. 2002 Mar;1(2):203-9.

1 In the Taskbar, select **Home**.

2 In the Home Page, click on the **Start An Analysis** link.

3 In the Choose An Analysis Workflow page, click on the **Gene Expression Profile(s)** link.

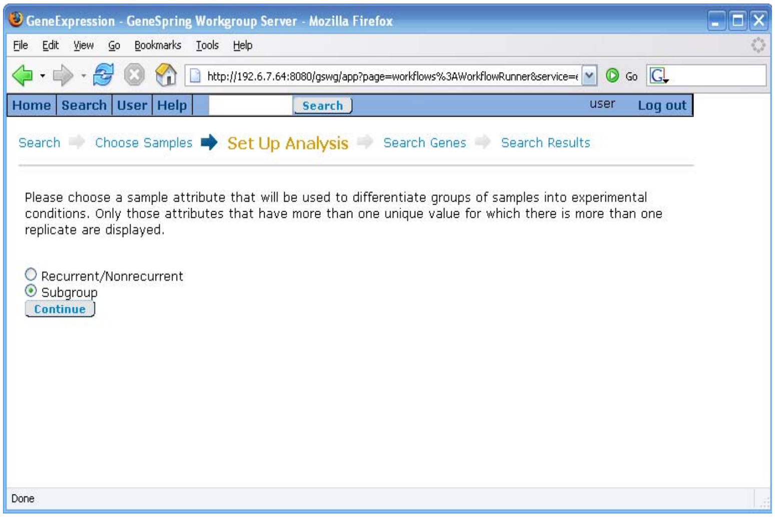
4 Search for samples using keywords or attributes (see “Searching and Inspecting Data” on page 20).

5 Select samples for analysis by checking the boxes in the first column of the tabular list that correspond to genes of interest (see “Data Selection” on page 27).

6 Click **Choose Samples** to analyze the samples selected in the previous step.

7 In the Set Up Analysis webform, select a comparison based on available attributes that will be used to differentiate groups of samples into experimental conditions. Only those attributes that have more than one unique value and for which there is more than one replicate are displayed.

Using the example data set, you select a differentiation attribute called subgroup, which corresponds to the surgical margin recorded following patient prostatectomy. Note that the Filter Genes option is checked by default.


GeneExpression - GeneSpring Workgroup Server - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.6.7.64:8080/gswg/app?page=workflows%3AWorkflowRunner&service=
Go
Home Search User Help
Search
user
Log out
Search
Choose Samples
Set Up Analysis
Search Genes
Search Results
Continue
Done

Please choose a sample attribute that will be used to differentiate groups of samples into experimental conditions. Only those attributes that have more than one unique value for which there is more than one replicate are displayed.

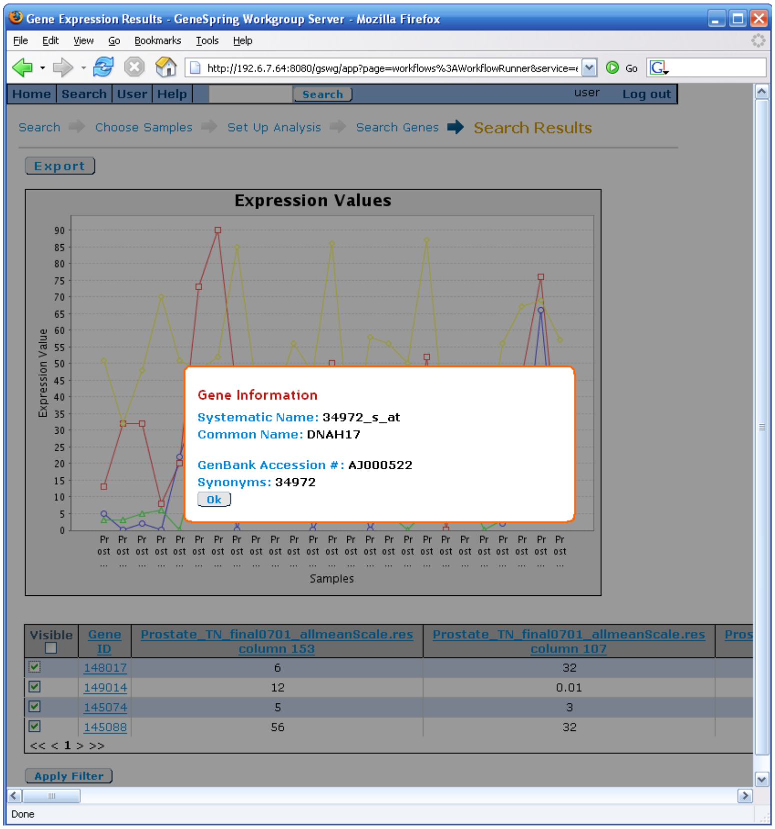
○ Recurrent/Nonrecurrent
● Subgroup

Figure 8 Set Up Analysis webform

8 Search the expression data for genes of interest (see “Search Genes Based on Nomenclature” on page 36).

9 Analyze and export the gene expression profiles visualization and table (see Figure 9 on page 19 and “Data Export” on page 37).

The lower portion of the Analysis Interface displays a tabular representation of the genes expression value across samples. Note that familiar table operations are available such as sorting and selection for filtering. Additional information is available through the Gene Information dialog box, shown layered on the workflow result page. The Gene Information dialog box can be displayed for any gene by clicking on a corresponding row name in the Gene ID column.



Gene Information

Systematic Name: 34972_s_at
Common Name: DNAH17

GenBank Accession #: AJ000522
Synonyms: 34972

Ok

Visible □	Gene ID	Prostate_TN_final0701_allmeanScale.res column 153	Prostate_TN_final0701_allmeanScale.res column 107	Pros
☑	148017	6	32	
☑	149014	12	0.01	
☑	145074	5	3	
☑	145088	56	32	

<< < 1 > >>

Figure 9 Gene Expression Profile(s) workflow result.

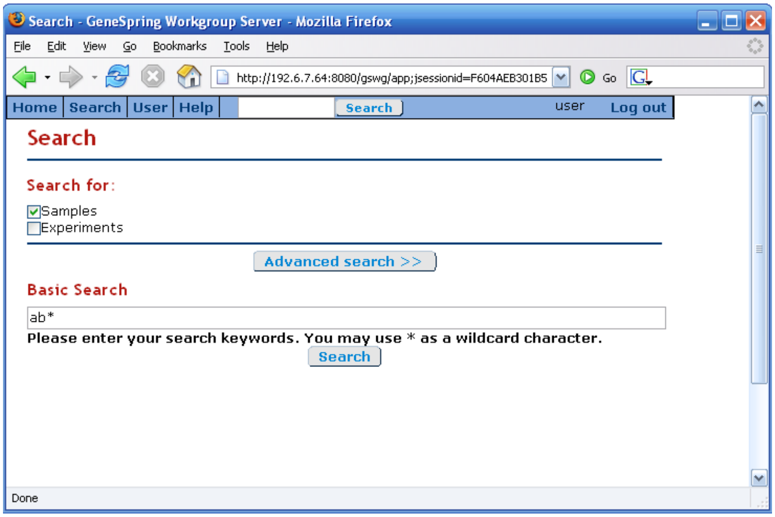
Searching and Inspecting Data

The Workgroup Server stores data from one and two-channel microarray experiments. which can be accessed for further analysis. Additionally, information relating to the experimental conditions are stored, allowing complex queries upon such metadata.

An important concept in the Web Application concerns microarray data mining, or effectively searching for and combining samples or experiments with contrasting conditions for analysis. When this step is applied during an analysis workflow, the Web Application will combine selected samples into a virtual experiment for the purposes of the current session. This section details use of the Web Application basic and advanced search capabilities as well as data inspection and data file attachment uploads to the database.

Basic Search

Upon selecting **Search** from the Taskbar or initiating a workflow analysis, the Web Application defaults to the basic search function, the output of which is identical to using the **Search Box** directly from within the Taskbar.


Search - GeneSpring Workgroup Server - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.6.7.64:8080/gswg/app;jsessionid=F604AEB301B5
Go
Home
Search
User
Help
Search
user
Log out
Advanced search >>
Search
Done

Search

Search for:

☑ Samples
□ Experiments

Basic Search

ab*

Please enter your search keywords. You may use * as a wildcard character.

Figure 10 The Basic Search webform

Based on specified search criteria, a list of samples or experiments will be returned, grouped by array type (referred to as a "genome" within both the Web Application and GeneSpring GX). The basic search function allows the following specifications to be made:

Search For By default, the Web Application will form a search of samples. If desired, Experiments (logical pre-ordered groupings of samples) can be searched in addition to or instead of samples alone. Selecting the appropriate checkbox toggles Sample and Experiment search targets.

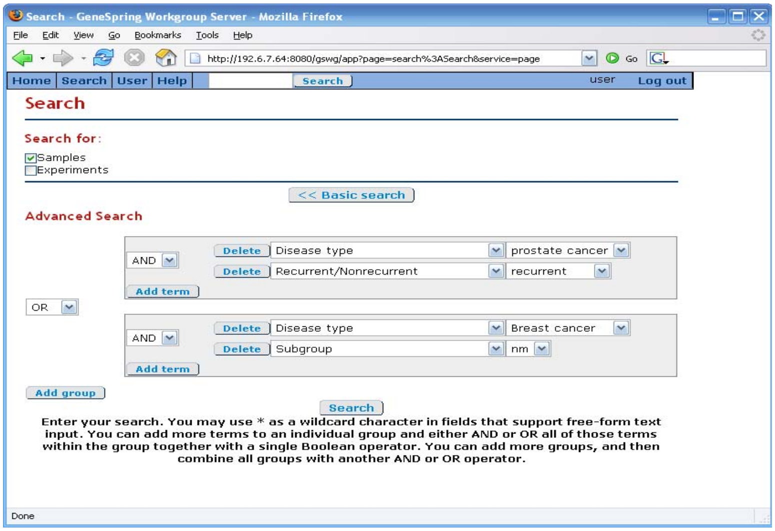
Advanced Search Clicking the **Advanced Search >>** button will navigate to a webform allowing additional criteria to be layered upon the basic search.

Basic Search Box A search of available microarray information by keywords entered into this box will occur. Use of the '*' character for wildcard searching is allowed. For example, to find samples, experiments, or both named aba, abb, and abc, you may use a search string "ab*". To retrieve all available samples, experiments, or both, you may use a pure wildcard search string "*".

Search Starts the basic search function.

Advanced Search

Advanced search may be accessed from the Basic Search webform by clicking on the **Advanced Search** button. Advanced search enables one or more criteria to be applied to the microarray information within Workgroup Server to attain specific search results.


Search - GeneSpring Workgroup Server - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.6.7.64:8080/gswg/app?page=search%3ASearch&service=page
Go
Home Search User Help
Search
user
Log out
<< Basic search
Delete
Add term
Add group
Search
Done

Search

Search for:

☑ Samples
☐ Experiments

Advanced Search

AND
Disease type: prostate cancer
Recurrent/Nonrecurrent: recurrent

OR

AND
Disease type: Breast cancer
Subgroup: nm

Enter your search. You may use * as a wildcard character in fields that support free-form text input. You can add more terms to an individual group and either AND or OR all of those terms within the group together with a single Boolean operator. You can add more groups, and then combine all groups with another AND or OR operator.

Figure 11 The Advanced Search webform.

Based on specified search criteria, a list of samples or experiments will be returned, grouped by array type (referred to as a “genome” within both the Web Application and GeneSpring GX). The advanced search function webform allows the following specifications to be made:

Search For

By default, the Web Application will search for samples. If desired, experiments (logical pre-ordered groupings of samples) can be searched in addition to or instead of samples alone. Selecting the appropriate checkbox toggles sample and experiment search targets.

Advanced Search

The advanced search webform is an expandable entry form that allows an arbitrary number of criteria to be combined. The subsections of the Advanced Search form consider both terms and groups of terms, each of which may be self-consistently applied using Boolean or set (AND/OR) operators.

Term Search Box

Web Application defaults to one term search, indicated by the gray bounding box. A drop down list consists of a logical set of microarray information along with any additional information uploaded into the Workgroup Server about available microarray samples and experiments. Upon selection of a search term, a second drop down box will appear if the selection term is part of a defined vocabulary. If the search term is not controlled, a text input box is presented. If no samples or experiments contain search term fields, the search term input will be removed from the form. The text input box accepts the '*' character as a wildcard.

The basic set of searchable microarray information terms include:

Table 1 Search Terms

Name	Information
• Accession ID	A unique identifier for each sample generated by Workgroup Server.
• Authors	The authors of the study.
• Attachment	Additional files associated with samples or experiments.
• Cache Timestamp	The timestamp that is used to regulate the cache.
• Date Created	The date that the samples were created in either GeneSpring GX or GeneSpring Workgroup Server.

Table 1 Search Terms (continued)

Name	Information
• Date To Release	If applicable, the date that the associated data may be made public.
• Date Uploaded	The date the data was loaded into Workgroup Server.
• Genome	The array design used.
• How Created	Information about the data file creation.
• Last Modified	Date that the sample or experiment was last modified.
• Name	Name of the sample or experiment.
• Note	Textual description of the sample or experiment.
• Organization	The name of the organization conducting the project.
• Original Source	The source of the biological material used in the microarray project.
• Owner	The Workgroup Server user responsible for the sample or experiment data.
• Parent Folder	The parent folder in the hierarchy of stored experiments. This search term does not apply to samples because they are stored in a flat directory structure within each genome.
• Previous Accession Ids	Identifiers previously used for the sample or experiment.
• Program Created	The program used to create the sample or experiment (for example, GeneSpring GX).
• Project	The name of the microarray project.
• Research Group	The name of the research group conducting the microarray project.

Table 1 Search Terms (continued)

Name	Information
• Date To Release	If applicable, the date that the associated data may be made public.
• Date Uploaded	The date the data was loaded into Workgroup Server.
• Genome	The array design used.
• How Created	Information about the data file creation.
• Last Modified	Date that the sample or experiment was last modified.
• Name	Name of the sample or experiment.
• Note	Textual description of the sample or experiment.
• Organization	The name of the organization conducting the project.
• Original Source	The source of the biological material used in the microarray project.
• Owner	The Workgroup Server user responsible for the sample or experiment data.
• Parent Folder	The parent folder in the hierarchy of stored experiments. This search term does not apply to samples because they are stored in a flat directory structure within each genome.
• Previous Accession Ids	Identifiers previously used for the sample or experiment.
• Program Created	The program used to create the sample or experiment (for example, GeneSpring GX).
• Project	The name of the microarray project.
• Research Group	The name of the research group conducting the microarray project.

Table 1 Search Terms (continued)

Name	Information
• Steps from Experimental Data	A GeneSpring GX generated field. For samples and experiments the value will always be zero.
• Technology	The software used to create the files loaded into Workgroup Server.
• Uploaded By	The name of the user who loaded the data into Workgroup Server.

(Term) AND/OR A drop down box within the gray bounding box of each set of search terms. This allows logical combinations of Terms in the search.

Add Term Allows additional search terms if required.

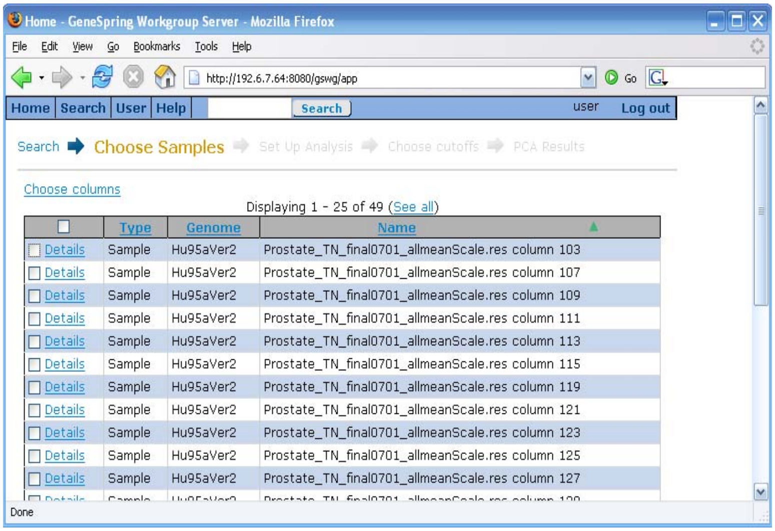
(Group) AND/OR A drop down box between the gray bounding boxes of each set of search terms (called a Group). This allows logical combinations of Groups in the search.

Add Group Clicking the **Advanced Search** button create an additional bounding box which may include layered search term(s). Each gray bounding box represents a group. Groups may be combined by set operators AND or OR available in a drop down box between each grouping.

Search Calls the search function using specified Terms and Groups.

Data Selection

Workgroup Web Application allows specific selection of information relating to a microarray samples or experiments following a search of available microarray data. The data is presented in tabular format.


Home - GeneSpring Workgroup Server - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.6.7.64:8080/gswg/app
Go
Home Search User Help
Search
user
Log out
Search
Choose Samples
Set Up Analysis
Choose cutoffs
PCA Results
Choose columns
Done

Displaying 1 - 25 of 49 (See all)

□	Type	Genome	Name ▲
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 103
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 107
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 109
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 111
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 113
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 115
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 119
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 121
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 123
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 125
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 127
□ Details	Sample	Hu95aVer2	Prostate_TN_final0701_allmeanScale.res column 129

Figure 12 The Choose Samples webform.

See all Selects all returned samples for display.

Column Headings Clicking on any bold blue column headings will sort the tabular search results according to that column header.

Sample Checkbox The first column of the tabular display contains a check box to include or exclude search result samples in further analysis. The check boxes in the header row will toggle selection or deselection of all samples shown on the page.

NOTE If the search returns multiple pages, the sample check box in the header row will only select those samples displayed on the page. For inclusion of all samples spanning multiple pages, click the See All link.

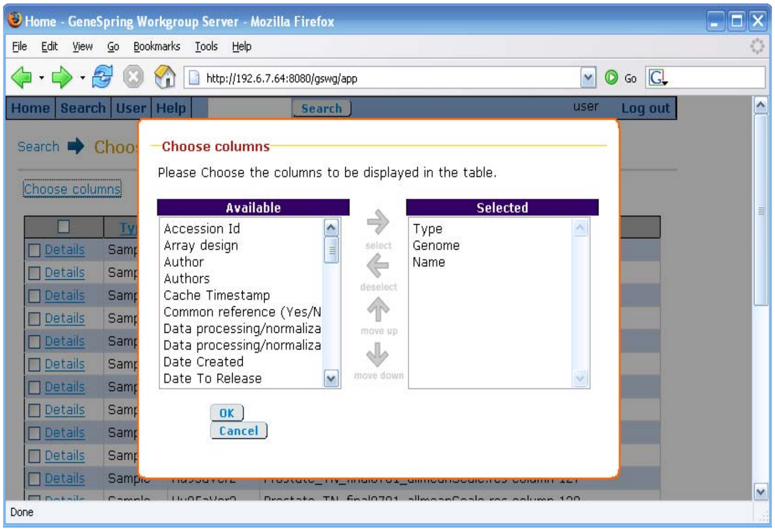
Details Each row returned by either the basic or advanced search contains a hyperlink named “Details” in the first column. Clicking on **Details** will navigate to a summary page for either sample inspection or experiment inspection.

<< < 1 2 > >> Paginates through returned samples.

Choose Samples Navigates forward one step in the analysis workflow using the selected samples. Only available if the samples or experiments are the result of a search within a workflow.

Choose Columns Clicking on the Choose columns link will display the Choose Columns dialog box which allows for selection of attributes to be included or excluded from the tabular display. The Choose Columns dialog box consists of two lists. The list titled Available displays fields that may be chosen to be shown in the tabular display. The list titled Selected displays those fields currently displayed in the tabular view. The Choose Columns dialog box has the following functions.

- The right hand arrow labeled Select moves user chosen fields from the Available list to the Selected list. Alternatively, double-click on a field in the Available list to move that field to the Selected list.
- The left hand arrow labeled Deselect moves user chosen fields from the Selected list to the Available list.
- The up arrow labeled Move Up moves a user chosen field to a higher level in the order of the list. This list order will be reflected in the order of columns in the tabular view.
- The down arrow labeled Move Down moves a user chosen field to a lower level in the order of the list. This list order will be reflected in the order of columns in the tabular view.



Home - GeneSpring Workgroup Server - Mozilla Firefox

http://192.6.7.64:8080/gswg/app

Choose columns

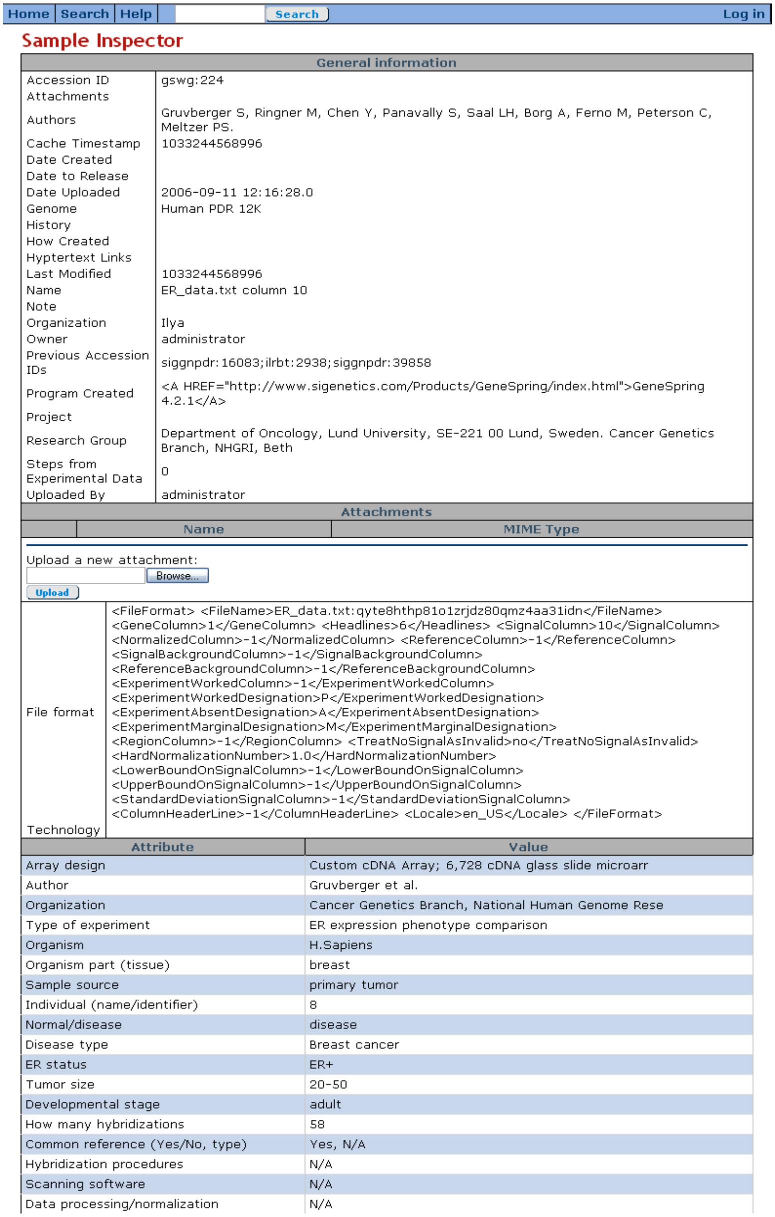
Please Choose the columns to be displayed in the table.

Available	Selected
Accession Id	Type
Array design	Genome
Author	Name
Authors	
Cache Timestamp	
Common reference (Yes/N	
Data processing/normaliza	
Data processing/normaliza	
Date Created	
Date To Release	

Figure 13 The Choose Columns dialog box (foreground).

Data Inspection

Workgroup Web Application allows detailed inspection of information relating to a microarray samples or experiments following a search of available microarray data. The data is available from the tabular views shown in “Data Selection” on page 27 by clicking on the Details link in each sample row.



Sample Inspector

General information	
Accession ID	gswg:224
Attachments	
Authors	Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS.
Cache Timestamp	1033244568996
Date Created	
Date to Release	
Date Uploaded	2006-09-11 12:16:28.0
Genome	Human PDR 12K
History	
How Created	
Hyptertext Links	
Last Modified	1033244568996
Name	ER_data.txt column 10
Note	
Organization	Ilya
Owner	administrator
Previous Accession IDs	siggnpdr:16083;ilrbt:2938;siggnpdr:39858
Program Created	<A HREF="http://www.sigenetics.com/Products/GeneSpring/index.html">GeneSpring 4.2.1</A>
Project	
Research Group	Department of Oncology, Lund University, SE-221 00 Lund, Sweden. Cancer Genetics Branch, NHGRI, Beth
Steps from Experimental Data	0
Uploaded By	administrator

Attachments

	Name	MIME Type

Upload a new attachment:

File format	<FileFormat> <FileName>ER_data.txt:qyte8hthp81o1zrjdz80qmz4aa31idn</FileName> <GeneColumn>1</GeneColumn> <Headlines>6</Headlines> <SignalColumn>10</SignalColumn> <NormalizedColumn>-1</NormalizedColumn> <ReferenceColumn>-1</ReferenceColumn> <SignalBackgroundColumn>-1</SignalBackgroundColumn> <ReferenceBackgroundColumn>-1</ReferenceBackgroundColumn> <ExperimentWorkedColumn>-1</ExperimentWorkedColumn> <ExperimentWorkedDesignation>P</ExperimentWorkedDesignation> <ExperimentAbsentDesignation>A</ExperimentAbsentDesignation> <ExperimentMarginalDesignation>M</ExperimentMarginalDesignation> <RegionColumn>-1</RegionColumn> <TreatNoSignalAsInvalid>no</TreatNoSignalAsInvalid> <HardNormalizationNumber>1.0</HardNormalizationNumber> <LowerBoundOnSignalColumn>-1</LowerBoundOnSignalColumn> <UpperBoundOnSignalColumn>-1</UpperBoundOnSignalColumn> <StandardDeviationSignalColumn>-1</StandardDeviationSignalColumn> <ColumnHeaderLine>-1</ColumnHeaderLine> <Locale>en_US</Locale> </FileFormat>
Technology	

Attribute	Value
Array design	Custom cDNA Array; 6,728 cDNA glass slide microarr
Author	Gruvberger et al.
Organization	Cancer Genetics Branch, National Human Genome Rese
Type of experiment	ER expression phenotype comparison
Organism	H.Sapiens
Organism part (tissue)	breast
Sample source	primary tumor
Individual (name/identifier)	8
Normal/disease	disease
Disease type	Breast cancer
ER status	ER+
Tumor size	20-50
Developmental stage	adult
How many hybridizations	58
Common reference (Yes/No, type)	Yes, N/A
Hybridization procedures	N/A
Scanning software	N/A
Data processing/normalization	N/A

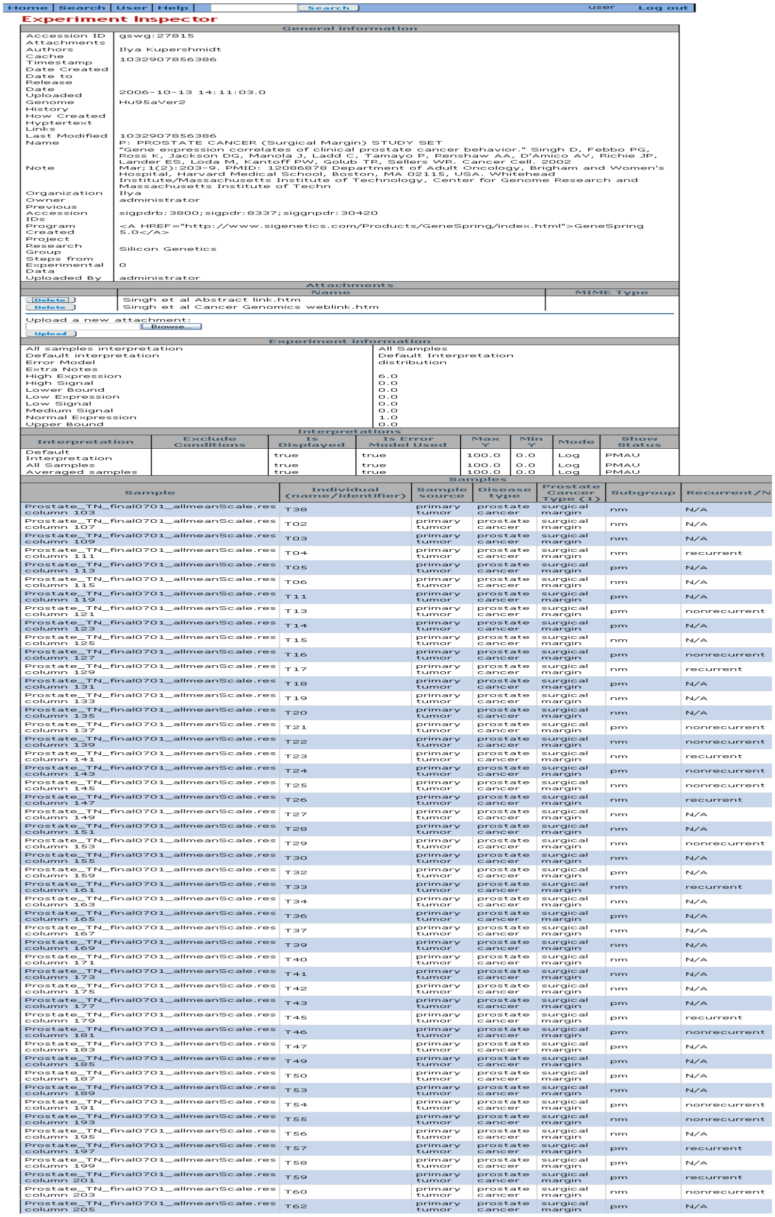
Figure 14 The Sample Inspector. Both the Sample Inspector and the Experiment Inspector allow for upload of associated files.

The Sample Inspector consists of three panels:

General Information Shows available history, administrative, and tracking information about the sample. The field values for a given sample will vary based upon the information attached to the sample in Workgroup Server.

Attachments This panel shows associated uploaded files and allows for additional files to be uploaded and associated with the sample data (for example, a pdf file of a related publication). The panel also shows the file format used to populate the Workgroup Server with sample information and the Software or Technology used to generate the data.

Attributes The bottom panel displays all available conditional information relating to the sample in a tabular view.



Experiment Inspector

General information

Accession ID	gswg:27815
Attachments	
Authors	Ilya Kupershmidt
Cache Timestamp	1032907856386
Date Created	
Date to Release	
Date Uploaded	2006-10-13 14:11:03.0
Genome	Hu95aVer2
History	
How Created	
Hyptertext Links	
Last Modified	1032907856386
Name	P: PROSTATE CANCER (Surgical Margin) STUDY SET
Note	"Gene expression correlates of clinical prostate cancer behavior." Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Cancer Cell. 2002 Mar;1(2):203-9. PMID: 12086878 Department of Adult Oncology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA. Whitehead Institute/Massachusetts Institute of Technology, Center for Genome Research and Massachusetts Institute of Techn
Organization	Ilya
Owner	administrator
Previous Accession IDs	sigpdrb:3800;sigpdr:8337;siggnpdr:30420
Program Created	<A HREF="http://www.sigenetics.com/Products/GeneSpring/index.html">GeneSpring 5.0</A>
Project	
Research Group	Silicon Genetics
Steps from Experimental Data	0
Uploaded By	administrator

Attachments

	Name	MIME Type
Delete	Singh et al Abstract link.htm	
Delete	Singh et al Cancer Genomics weblink.htm	

Upload a new attachment:

Browse...

Upload

Experiment information

All samples interpretation	All Samples
Default interpretation	Default Interpretation
Error Model	distribution
Extra Notes	
High Expression	6.0
High Signal	0.0
Lower Bound	0.0
Low Expression	0.0
Low Signal	0.0
Medium Signal	0.0
Normal Expression	1.0
Upper Bound	0.0

Interpretations

Interpretation	Exclude Conditions	Is Displayed	Is Error Model Used	Max Y	Min Y	Mode	Show Status
Default Interpretation		true	true	100.0	0.0	Log	PMAU
All Samples		true	true	100.0	0.0	Log	PMAU
Averaged samples		true	true	100.0	0.0	Log	PMAU

Samples

Sample	Individual (name/identifier)	Sample source	Disease type	Prostate Cancer Type (1)	Subgroup	Recurrent/N
Prostate_TN_final0701_allmeanScale.res column 103	T38	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 107	T02	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 109	T03	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 111	T04	primary tumor	prostate cancer	surgical margin	nm	recurrent
Prostate_TN_final0701_allmeanScale.res column 113	T05	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 115	T06	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 119	T11	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 121	T13	primary tumor	prostate cancer	surgical margin	pm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 123	T14	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 125	T15	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 127	T16	primary tumor	prostate cancer	surgical margin	pm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 129	T17	primary tumor	prostate cancer	surgical margin	nm	recurrent
Prostate_TN_final0701_allmeanScale.res column 131	T18	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 133	T19	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 135	T20	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 137	T21	primary tumor	prostate cancer	surgical margin	pm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 139	T22	primary tumor	prostate cancer	surgical margin	nm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 141	T23	primary tumor	prostate cancer	surgical margin	nm	recurrent
Prostate_TN_final0701_allmeanScale.res column 143	T24	primary tumor	prostate cancer	surgical margin	pm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 145	T25	primary tumor	prostate cancer	surgical margin	nm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 147	T26	primary tumor	prostate cancer	surgical margin	nm	recurrent
Prostate_TN_final0701_allmeanScale.res column 149	T27	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 151	T28	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 153	T29	primary tumor	prostate cancer	surgical margin	nm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 155	T30	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 159	T32	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 161	T33	primary tumor	prostate cancer	surgical margin	nm	recurrent
Prostate_TN_final0701_allmeanScale.res column 163	T34	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 165	T36	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 167	T37	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 169	T39	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 171	T40	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 173	T41	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 175	T42	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 177	T43	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 179	T45	primary tumor	prostate cancer	surgical margin	pm	recurrent
Prostate_TN_final0701_allmeanScale.res column 181	T46	primary tumor	prostate cancer	surgical margin	pm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 183	T47	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 185	T49	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 187	T50	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 189	T53	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 191	T54	primary tumor	prostate cancer	surgical margin	pm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 193	T55	primary tumor	prostate cancer	surgical margin	nm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 195	T56	primary tumor	prostate cancer	surgical margin	nm	N/A
Prostate_TN_final0701_allmeanScale.res column 197	T57	primary tumor	prostate cancer	surgical margin	pm	recurrent
Prostate_TN_final0701_allmeanScale.res column 199	T58	primary tumor	prostate cancer	surgical margin	pm	N/A
Prostate_TN_final0701_allmeanScale.res column 201	T59	primary tumor	prostate cancer	surgical margin	pm	recurrent
Prostate_TN_final0701_allmeanScale.res column 203	T60	primary tumor	prostate cancer	surgical margin	nm	nonrecurrent
Prostate_TN_final0701_allmeanScale.res column 205	T62	primary tumor	prostate cancer	surgical margin	pm	N/A

Figure 15 The Experiment Inspector. Both the Sample Inspector and the Experiment Inspector allow for upload of associated files.

The Experiment Inspector consists of five panels:

General Information Shows available history, administrative, and tracking information about the experiment. The field values for a given experiment will vary based upon the information attached to the experiment in Workgroup Server.

Attachments This panel shows associated uploaded files and allows for additional files to be uploaded and associated with the experiment data (for example, a PDF file of a related publication).

Experiment Information The experiment information panel displays aggregate summary information for all samples within an experiment including interpretation modes, error models, and summary statistics.

Interpretations This panel summarizes the interpretations applied to the experiment and the parameters of each interpretation.

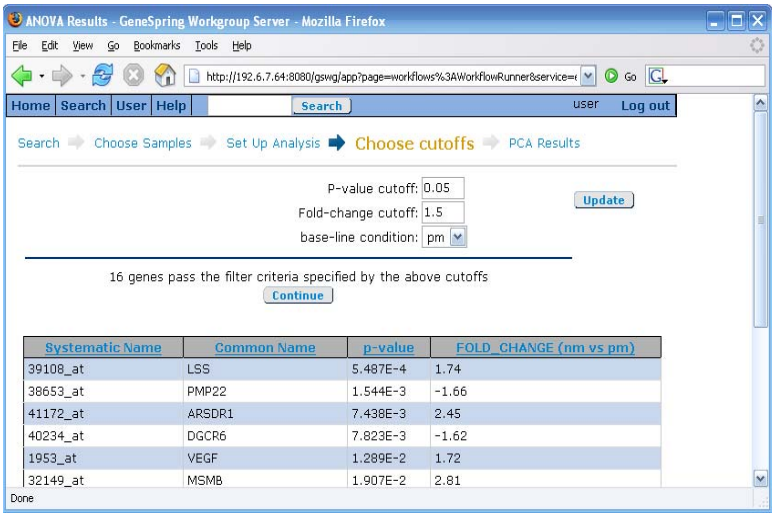
Samples The bottom panel displays in tabular form each sample considered in the experiment, the individual attributes, and the sample source specific to the experiment.

Gene Selection

Analysis workflow proceeds from selected samples returned by specified search criteria. For PCA and Heatmap generation workflows, an additional step is available to filter genes based upon thresholds imposed upon p-value and fold-change scores. For Gene Expression Profile(s) analysis, an additional step is necessary to select genes for examination based upon keyword values.

Choose Cutoffs for Gene Filtering Based on Scores

The Analysis Interface will display the current number of genes that pass filtering and a tabular view of the gene systematic name, common name, p-value, and fold-change. For information on how p-values are calculated, see “p-value determination” on page 40.


ANOVA Results - GeneSpring Workgroup Server - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.6.7.64:8080/gswg/app?page=workflows%3AWorkflowRunner&service=
Go
Home Search User Help
Search
user
Log out
Search
Choose Samples
Set Up Analysis
Choose cutoffs
PCA Results
Update
Continue
Done

P-value cutoff: 0.05

Fold-change cutoff: 1.5

base-line condition: pm

16 genes pass the filter criteria specified by the above cutoffs

Systematic Name	Common Name	p-value	FOLD_CHANGE (nm vs pm)
39108_at	LSS	5.487E-4	1.74
38653_at	PMP22	1.544E-3	-1.66
41172_at	ARSDR1	7.438E-3	2.45
40234_at	DGCR6	7.823E-3	-1.62
1953_at	VEGF	1.289E-2	1.72
32149_at	MSMB	1.907E-2	2.81

Figure 16 The Choose Cutoffs webform for PCA and Heatmap workflow gene filtering.

P-value cutoff Statistical analysis is automatically performed on the experimental conditions specified in the Set Up Analysis webform. If two conditions are being compared, Welch’s t-test is performed to identify differentially expressed genes. If three or more conditions are being compared, Welch’s ANOVA is performed. No multiple testing correction is applied. To filter genes based on their associated p-values, you must specify a p-value cutoff. If you defined the p-value cutoff to be 0.05, genes with p-values less than or equal to 0.05 (<=0.05), as calculated by the statistical test, will be displayed in the Filter Genes webform.

Fold-change cutoff To filter genes based on fold-change, you must specify both the condition to set as the baseline for fold-change and the minimum fold-change threshold that a gene must pass in at least one comparison for it to pass the filter. For example, suppose the conditions tested are “control treatment” and “A”, “B”, and “C” treatments. If the baseline condition is set to “control treatment” and the minimum fold-change is set to 3, genes with expression values greater than (>3) three fold in magnitude in at least one of the three paired comparisons will pass the filter and be displayed.

Base-line condition The condition to be used as a common base for comparisons when determining fold-changes. Note that in the case of any two comparisons, the selection yields symmetric results. In other words, a fold-change of 2 in a comparison of attribute A vs. base-line attribute B will result in a fold-change of -2 in a comparison of A vs. base-line attribute B. Note that fold change values of both +2 and -2 will pass a cutoff value of any positive number less than +2 or any negative number greater than -2.

Update Refreshes the tabular gene list that passes filtering conditions.

Continue Proceeds to the final analysis step for either PCA or Heatmap analysis.

<< < 1 2 > >> Scrolls through the display pages of the filtered gene list.

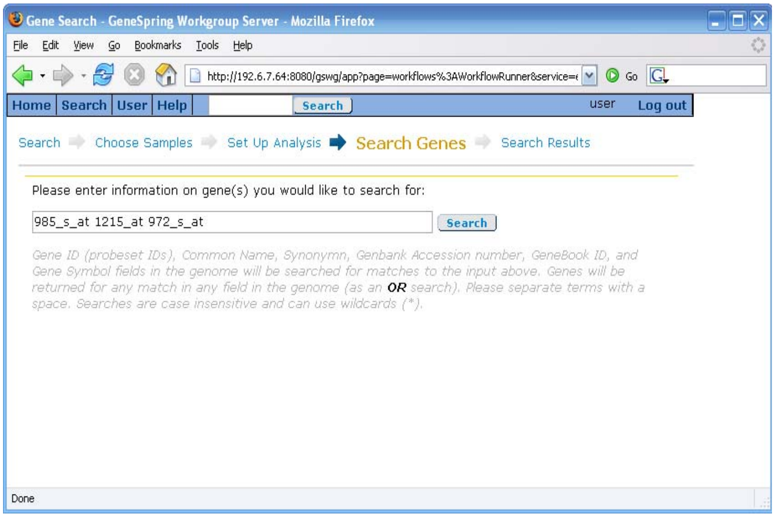
Column Headings Clicking on any bold blue column headings will sort the tabular gene filtering results according to that column header.

Search Genes Based on Nomenclature

To view the expression data for genes of interest, search for the genes using any of the following search terms:

- Gene ID (Probeset IDs)
- Common Name
- Synonym
- Genbank Accession Number
- GeneBook ID
- Gene Symbols

Multiple genes can be searched simultaneously by entering search terms for each gene delimited by white space. The search will proceed if any of the search strings are met (as though each term is separated by a logical OR statement). Search strings are case sensitive and allow the use of the '*' character as a wildcard.



http://192.6.7.64:8080/gswg/app?page=workflows%3AWorkflowRunner&service=

Please enter information on gene(s) you would like to search for:

985_s_at 1215_at 972_s_at

Gene ID (probeset IDs), Common Name, Synonymn, Genbank Accession number, GeneBook ID, and Gene Symbol fields in the genome will be searched for matches to the input above. Genes will be returned for any match in any field in the genome (as an ***OR*** *search). Please separate terms with a space. Searches are case insensitive and can use wildcards (*).*

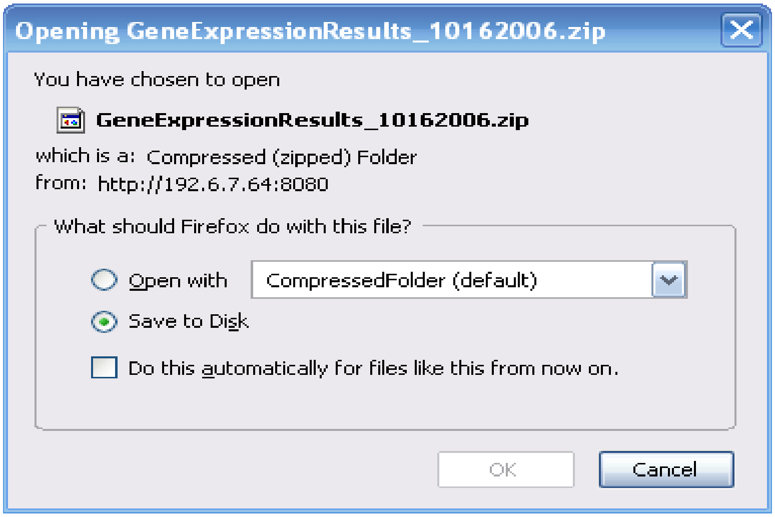
Figure 17 The Search Genes webform for gene selection prior to gene expression profile analysis.

Data Export

Analysis workflow results can be exported for further analysis. Data export creates a compressed zip folder for download from the Web Application.

Data may be exported from the results page of any analysis workflow, which displays an **Export** button at the top of the analysis interface.

Export Choose to save the created folder to your desktop and then access it using an available zip extraction program.



Opening GeneExpressionResults_10162006.zip

You have chosen to open

GeneExpressionResults_10162006.zip

which is a: Compressed (zipped) Folder

from: http://192.6.7.64:8080

What should Firefox do with this file?

○ Open with CompressedFolder (default)

● Save to Disk

□ Do this automatically for files like this from now on.

Figure 18 Data export creates a compressed zip folder. Choose to save this folder to disk.

The contents of the zip folders are dependent upon the workflow.

Sample Comparison (PCA) The PCA analysis workflow will generate fixed scatterplot image files and a tabular text file. In all scatterplots, the points are colored using the chosen parameters. Axes are labeled with the percent explained variance for the corresponding principle component (as in GeneSpring GX).

- A fixed 3-D scatterplot png file showing samples plotted using PCA scores for the first three principal components.
- Three fixed 2-D scatterplot .png files showing samples plotted using PCA scores for the two principle components (one for each paired comparison of the three principle components).
- A table text file containing scores for each sample against the first three principal components (i.e. the coordinates used in the 3-D scatterplot).

Differential Expression with Heatmap

The Heatmap analysis workflow generates a static png image of the gene tree and condition tree with the following features.

- Branches in the condition tree are colored by the chosen parameters.
- Coloring blocks are shown for the chosen parameters.

Gene Expression Profile(s)

The Gene Expression Profile(s) analysis workflow generates a single fixed-line graph png file displaying the expression profiles of all of the selected genes across the selected set of samples. A tabular text file is also generated with the following columns:

- Systematic Name.
- Common Name.
- GenBank Accession Number.
- Synonyms.

Statistical References

GeneSpring Workgroup Web Application performs basic exploratory analysis on stored microarray data using PCA, Heatmap visualization, and gene expression profile comparisons as well as preprocessing and p-value generation. The underlying functionality is derived from Agilent GeneSpring GX 7.3 and utilizes default settings in that application.

The following sections briefly describe those statistical measures and their default parameters. Each algorithm is well described in microarray analysis literature and the interested reader may benefit from a more thorough examination of the underlying mathematics behind such analysis.

Normalization and Pre-Processing

RMA normalization is performed on Affymetrix samples upon loading into GeneSpring Workgroup Server by the Workgroup SampleLoader after samples are selected in the Web Application. A GeneSpring virtual experiment is then created from the samples with the following settings:

1. GeneSpring GX normalizations: none for Affymetrix samples; Lowess normalization as necessary for 2-color samples.
2. Interpretation: Data is processed in log base 10 mode. In GeneSpring GX this is referred to as a log interpretation.
3. Gene list: all genes will be used for analysis unless otherwise specified in the workflows.

RMA normalization

Robust Multichip Averaging (RMA) is a normalization procedure consisting of three steps:

1. Background correction using Affymetrix PM probe information. In RMA, MM probes are discarded for use in estimating background noise and each array is assumed to have a common mean background.
2. Normalization across all chips in a given set. Quantile normalization is applied and the intensities are adjusted to

produce similar distributions. This is accomplished by transforming the value of the quantile in the distribution of probe intensities to the quantile's value on the reference chip.

3 Expression measure summary. Median polish is used to estimate log expression.

Lowess normalization In two-channel microarrays, variations in intensities between the channels leads to an effect known as dye bias. Such a bias may introduce spurious results. Lowess normalization merges and smooths two-channel data on each sample in the set independently to reduce such effects. Lowess consists of:

1 A lowess fit (or locally weighted least squares regression) of a smooth curve to the data set produced by a log(ratio) vs. log(sqrt(product)) matrix (also known as an MA plot, or a plot which measures ratios by intensities).

2 An adjustment of the log(ratio) value by the lowess fit.

p-value determination

A p-value is a statistical measure that represents the probability of obtaining a result of the same magnitude or greater than the result obtained if the null hypothesis is true. Therefore, smaller p-values are indicative of greater belief in the reproducibility of a given test. To obtain p-values used in the filtering function of the PCA and Heatmap workflows, Welch's t-test is applied to the case of a two condition comparison and Welch's ANOVA is performed for three or more condition comparisons. Welch's ANOVA is performed corresponding to the following settings in GeneSpring GX:

1 No equal variances assumption.

2 p-value cutoff of 1 (all genes returned).

3 No multiple testing correction or post-hoc tests.

Welch's t-test The t-test assesses whether the means of two groups are significantly different. A simplified view of the t-test is to measure the difference in means and divide by the sample standard deviation normalized by the square root of the sample size. Welch's t-test negates the assumption of equal variances between the groups.

ANOVA The Analysis of Variance (ANOVA) is a statistical test that divides the total variability of the data into variability between and within groups. If the variability between groups is large compared to the variability within groups then the test will suggest a significant difference between the groups. Welch's ANOVA negates the assumption of equal variances between the groups.

Hierarchical Clustering

Dendrograms created for the Heatmap analysis are a method of hierarchical clustering and follow the GeneSpring GX 7.3 default settings:

1. All samples interpretation.
2. Pearson correlation for similarity.
3. Average linkage.

Pearson correlation The Pearson Product Moment Correlation reflects the degree of linear relationship between variables. It can be described as the covariance of the two variables normalized by their standard deviations. The Pearson correlation results in a score ranging from -1 for inversely correlated variables to +1 for correlated variables.

Average linkage In clustering, a decision must be made as to how to measure the distance (often in euclidian space) between any two groups, or putative clusters. Since clustering is always an iterative process of either adding a member to a putative cluster or subtracting a member from a putative cluster, small differences in the distance space between cluster subgroups can have profound results in the reassignment of a member to a cluster. In average linkage, the distance between any two clusters is calculated as the arithmetic mean of distances between all possible pairs from the two clusters. Average linkage is considered a compromise between methods that minimize or maximize the cluster distances.

Principle Component Analysis

Principle Component Analysis (PCA) is a linear transformation of data with high dimensionality to a data set with lower dimensionality where each reduced dimension is orthogonal, and therefore independent of each other. PCA aims to identify conditions that contribute to the variation of the data set when visualization is impossible because of the large number of dimensions that describe the data. It is commonly used in microarray applications because of the ability to visualize trends in such multivariate data sets.

A simplified view of dimensional reduction involves a projection; if you have an n-dimensional scatterplot and pretend that each point is a physical ball, then shining a light through the scatterplot at various angles would result in different projections, or views, of the data on a two-dimensional backdrop such as a movie screen. One appealing aspect of such an approach is that you can repeat the projection as often as you like from different angles, looking for sources of variation along the way. It would be advantageous to represent each source of variation in the data by a new dimension and visualize the resultant plot. Looking at the projection on the movie screen, a best fit line to the projection that yields the greatest spread in the data may be applied. For instance, if the cloud of data points is shaped like a football, the main direction of the data would be a midline or axis along the length of the football. This also means that this line explains best the patterns observed in the data set. This line is known as the first principle component and will form the basis of a new coordinate system describing our data. This is the basis of a linear transformation.

Projections with small angle deviations from the first principle component are well explained by the principle component. Projections with large angle deviations from the first principle component indicates that the first principle component has little influence on the pattern of the rest of the data. Since it is desirable to describe the data in terms of factors that have a large influence, the next step is to choose an orthogonal projection which yields the (second) greatest variation in the data. This is known as the second principle component. Repeating this projection as often as necessary (always from orthogonal directions) the data set can be transformed into a new coordinate system. The number of projections may be limited to a manageable number (most data is explained very well by three principle components and subsequent projections yield little additional knowledge). Such reduction of dimensionality, or decomposition, is now on a coordinate system known by its eigenvectors with values known as PCA scores.

Dimensional reduction allows you to arbitrarily explore underlying relationships in the data. For example, even a 2-dimensional spreadsheet (which is actually a matrix, with one vector in each the horizontal and vertical directions) can be difficult to analyze. An understanding of the relationships and sources of variation in a very large 2 dimensional matrix is daunting. However, a simple transformation allows you to plot the data on a graph, where each column and row of the matrix is a vector in cartesian coordinates. The resultant plots show variation visually. It is possible to do such an analysis in a spreadsheet application, but it becomes challenging to analyze beyond three dimensions.

Interpreting PCA results is more challenging than applying the transformation. A common strategy is to take pairwise comparisons of the first three principle components, yielding clear views of the data from all sides. In microarray analysis, PCA is used for understanding variability and classification model testing. For example, if a classification scheme is to be applied to a data set, that classification scheme should, in principle, be markedly delineated along the principle components. Repeated selection of genes indicated in the classification model can allow for model parameter estimation in classification building or gene identification given a model.

Although small degrees of rotation around the principle component yield projections well explained by the principle component, PCA is not a considered clustering tool. The goal of PCA is to effectively ‘summarize’ the data and as such it is most effectively a visualization technique.

www.agilent.com

In this book

The *User Guide* provides information on the use of GeneSpring Workgroup 5.3 Web Application for mining and analysis of microarray data within Workgroup Server.

First Edition, October 2006

Agilent Technologies